**Introducción**

De acuerdo con la ONU [1], en el mundo hay más de 700 millones de personas que viven en situación de extrema pobreza, comprendida por la incapacidad de satisfacer las necesidades básicas, como son la salud, la educación, el acceso al agua, entre otras. Uno de los parámetros para identificar a las personas que viven en estas condiciones es la pobreza monetaria, que es estimado indirectamente a través de la comparación del ingreso per cápita con el costo monetario de adquirir una canasta de bienes [2]. En Colombia, desde el 2002 la tendencia de la pobreza monetaria ha sido decreciente, sin embargo, además de que la tasa de reducción se ha disminuido, el valor al 2018 era del 27% [2]. De acuerdo con el Banco Mundial [3], las zonas con mayor pobreza en Colombia a 2015 eran el Choco, La Guajira y el Cauca con valores de 26.1%, 20.9% y 12.4%, respectivamente.

La predicción de la pobreza es un objetivo crítico para lograr la implementación efectiva de políticas, que permita direccionar los recursos eficientemente a aquellos hogares que están en una situación de pobreza o que están a punto de llegar a esa situación [4]. Dado que los censos nacionales son costosos y complejos logísticamente y que los modelos económicos utilizados hasta el momento dependen de varias suposiciones, los métodos de *machine learning* se ven como una opción prometedora, especialmente cuando la cantidad de información ha aumentado considerablemente en los últimos años [5].

En este trabajo, con el objetivo de clasificar una muestra de hogares en Colombia de si son pobres o no, se presentan algunos métodos de *machine learning* utilizados para resolver problemas de clasificación y los resultados obtenidos, tomando como fuente de datos la base del DANE correspondiente a la *Medición de Pobreza Monetaria y Desigualdad 2018* en Colombia. Estos datos corresponden a encuestas realizadas a hogares del país en la Gran Encuesta Integrada de Hogares - GEIH 2018.

Para este ejercicio se utilizaron cuatro bases de datos, dos correspondientes a muestras de los hogares, una de entrenamiento y otra de prueba, con variables físicas del hogar, del arriendo, de los ingresos por unidad de gasto y de la clasificación realizada por el DANE de la pobreza monetaria (estas dos últimas variables solamente están en la de entrenamiento), y las otras dos correspondientes a individuos dentro de estos hogares, igualmente una de entrenamiento y otra de prueba, con variables sociodemográficas, de aspectos laborales (tipo de ocupación, recibo de subsidios o de bonificaciones, entre otras) y de varios tipos de ingreso (estos últimos solamente en la base de entrenamiento).

Resultados y conclusiones resumidas

**Datos**

Las bases de datos utilizadas para el ejercicio contienen varias variables de interés que están identificadas con la nomenclatura utilizada por la DIAN. Luego de identificar el tipo de variable y la información contenida, se renombraron aquellas que se consideraron como importantes para la predicción correspondiente. Dado que el enfoque de este ejercicio requiere predecir si un hogar es pobre o no mediante dos caminos, el primero siendo la predicción directa de la variable *Pobre* y el segundo la predicción del ingreso del hogar, la selección de estas variables debía contemplar aquellas características de los hogares que pudieran influir en esta clasificación, por ejemplo, el lugar en el que están ubicados, el número de habitaciones, tanto habitadas como no habitadas, el tipo de propiedad (arrendada, propia, etc.), las variables *proxy* del gasto realizado que demanda la propiedad y la cantidad de personas que viven allí; por otro lado, para los individuos, dado que la clasificación de los hogares en términos de pobreza depende del ingreso que conjuntamente se genera en el hogar por los individuos que viven allí, es fundamental mantener las variables que se consideran necesarias para predecir adecuadamente el ingreso de la persona.

A partir de la teoría económica y del primer taller, para las bases de datos de los individuos, se consideraron la educación, el sexo y la edad; adicional a estas, teniendo en consideración que el tipo de ocupación, la formalidad del trabajo, el hecho de recibir subsidios y bonificaciones, el estado laboral, tienen influencia sobre los ingresos, se incluyeron dentro del análisis.

Por otro lado, teniendo en cuenta que desde las bases de datos de los hogares se debe predecir la condición de pobreza y que esta depende del ingreso de los individuos, deben crearse nuevas variables que agrupen características relevantes para la determinación del ingreso; por ejemplo, el sexo de la jefe de hogar o el porcentaje de personas inactivas dentro del hogar, entre otros, pueden tener relevancia en que el hogar pueda tener más o menos ingresos.

**Bibliografía**

[1] Organización de las Naciones Unidas (2023). Objetivo 1: Poner fin a la pobreza en todas sus formas en todo el mundo. <https://www.un.org/sustainabledevelopment/es/poverty/>

[2] Departamento Nacional de Planeación (2019). Pobreza monetaria y pobreza multidimensional. Análisis 2008 – 2018. <https://colaboracion.dnp.gov.co/CDT/Desarrollo%20Social/Documento%20de%20An%C3%A1lisis%20de%20las%20Cifras%20de%20Pobreza%202018.pdf>

[3] Banco Mundial (2019). Global Subnational Atlas of Poverty (GSAP). <https://pipmaps.worldbank.org/en/data/datatopics/poverty-portal/poverty-geospatial?dataset=PovertyRate2.15-gsap&zoomLevel=7&lat=14.195163013871356&lng=-86.45141601562501>

[4] Qing Li, Shuai Yu, Damien Échevin, Min Fan (2022). Is poverty predictable with machine learning? A study of DHS data from Kyrgyzstan, Socio-Economic Planning Sciences, Volume 81, 2022, 101195, ISSN 0038-0121, <https://doi.org/10.1016/j.seps.2021.101195>.

[5] Pathways for Prosperity Commission (2023). Can machine learning predict poverty? <https://pathwayscommission.bsg.ox.ac.uk/blog/can-machine-learning-predict-poverty/>